Tuning a Kernel Designed to Process Protein Sequences

Yann Guermeur

LORIA - CNRS

http://www.loria.fr/~guermeur/

Overview

Multi-class SVM for protein sequence processing

- M-SVMs
- Kernel for protein sequence processing
- Parameterization

Guaranteed risks for discriminant models : state-of-the-art

- Bi-class case : binary-valued functions
- Bi-class case : real-valued functions
- Multi-class case : $\{1, \ldots, Q\}$-valued functions

Guaranteed risks for multi-class discriminant models : our contribution

- Multi-class case : vector-valued functions
- Generalized VC dimensions
- Bound on the risk for M-SVMs

Statistical learning - Pattern recognition

Hypotheses : empirical data characterizing a joint probability distribution

- Q-category discrimination problem
- $Z = (X, Y)$: random variable on a probability space
- $X(\Omega) = \mathcal{X}$: input space (set of *descriptions*), $Y(\Omega) = \mathcal{Y}$: finite set of *categories*
- F : joint probability distribution function on $\mathcal{X} \times \mathcal{Y}$, *fixed but unknown*
- $s = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$, *learning set* : observations i.i.d. according to F
- $\mathcal{H}$: family of vector-valued functions $h = [h_k]$, $(1 \leq k \leq Q)$, from $\mathcal{X}$ into $\mathbb{R}^Q$

Goal : for a given pattern, find its category

Find in $\mathcal{H}$ a function associated with the lowest *expected risk (generalization error)*

$$R(h) = R(f) = \mathbb{P}\{(x, y) : \ f(x) \neq y\} = \int_{\mathcal{X} \times \mathcal{Y}} \mathbb{1}_{\{f(x) \neq y\}} dF(x, y)$$

f : discriminant function corresponding to h, obtained by choosing the category associated with the index of the highest output

Multi-class Support Vector Machines (M-SVMs)

Architecture (common to all machines)

Functions $h = [h_k]$ computed by the architecture are defined by :

$$\forall x \in \mathcal{X},\ \forall k \in \{1, \dots, Q\},\ h_k(x) = \tilde{h}_k(x) + b_k = \langle w_k, \Phi(x) \rangle + b_k$$

where

- Φ is either a nonlinear map into a *feature space* or identity
- $\tilde{\mathcal{H}} = \left\{ \tilde{h} = \left[\tilde{h}_k \right] \right\}$ is the product of Q *reproducing kernel Hilbert spaces* (RKHS)
- κ, the *kernel* associated with $\tilde{\mathcal{H}}$, is related to Φ through :

$$\forall (x^{(1)}, x^{(2)}) \in \mathcal{X}^2,\ \kappa(x^{(1)}, x^{(2)}) = \langle \Phi(x^{(1)}), \Phi(x^{(2)}) \rangle$$

Training algorithm of M-SVMs : general principle

Let $s_m = ((x_1, C(x_1)), \ldots, (x_m, C(x_m)))$ be the training set

Primal formulation (with *soft margins*)

Problem 1

$$\min_{h \in \mathcal{H}} \left\{ \frac{1}{2} \|\tilde{h}\|^2_{\tilde{\mathcal{H}}} + C \sum_{i=1}^{m} l\left(h(x_i), C(x_i)\right) \right\}$$

Dual formulation

Representer theorems establish that training (solving Problem 1) amounts to finding the values of the coefficients β_{ik} in :

$$\forall k \in \{1, \ldots, Q\}, \; h_k(x) = \sum_{i=1}^{m} \beta_{ik} \kappa(x_i, x) + b_k$$

Training algorithms of M-SVMs (primal formulation)

Problem 2 (M-SVM1 (Vapnik & Blanz, 98, Weston & Watkins, 98,...))

$$\min_{h \in \mathcal{H}} \left\{ \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 + C \sum_{i=1}^{m} \sum_{k=1}^{Q} \xi_{ik} \right\}$$

$$s.t. \begin{cases} (w_{C(x_i)} - w_k)^T x_i + b_{C(x_i)} - b_k \geq 1 - \xi_{ik}, & (1 \leq i \leq m), (1 \leq k \neq C(x_i) \leq Q) \\ \xi_{ik} \geq 0, & (1 \leq i \leq m), (1 \leq k \neq C(x_i) \leq Q) \end{cases}$$

Problem 3 (M-SVM2 (Guermeur, 02))

$$\min_{h \in \mathcal{H}} \left\{ \frac{1}{2} t^2 + C \sum_{i=1}^{m} \sum_{k=1}^{Q} \xi_{ik} \right\}$$

$$s.t. \begin{cases} \|w_k - w_l\|^2 \leq t^2, \ (1 \leq k < l \leq Q) \\ \textit{Constraints of Problem 2} \end{cases}$$

Training algorithms of M-SVMs (primal formulation)

Problem 4 (M-SVM3 (Crammer & Singer, 01))

$$\min_{h \in \mathcal{H}} \left\{ \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 + C \sum_{i=1}^{m} \xi_i \right\}$$

$$s.t.\ \langle w_{C(x_i)} - w_k, \Phi(x_i) \rangle + b_{C(x_i)} - b_k + \delta_{C(x_i),k} \geq 1 - \xi_i,\ \ (1 \leq i \leq m), (1 \leq k \leq Q)$$

Problem 5 (M-SVM4 (Lee *et al.*, 01))

$$\min_{h \in \mathcal{H}} \left\{ \frac{1}{2} \sum_{k=1}^{Q} \|w_k\|^2 + C \sum_{i=1}^{m} \sum_{k=1}^{Q} \xi_{ik} \right\}$$

$$s.t. \begin{cases} \sum_{k=1}^{Q} w_k = 0,\ \ \sum_{k=1}^{Q} b_k = 0 & \\ \langle w_k, \Phi(x_i) \rangle + b_k \leq -1/(Q-1) + \xi_{ik}, & (1 \leq i \leq m), (1 \leq k \neq C(x_i) \leq Q) \\ \xi_{ik} \geq 0, & (1 \leq i \leq m), (1 \leq k \neq C(x_i) \leq Q) \end{cases}$$

Behaviour of a M-SVM (M-SVM1)

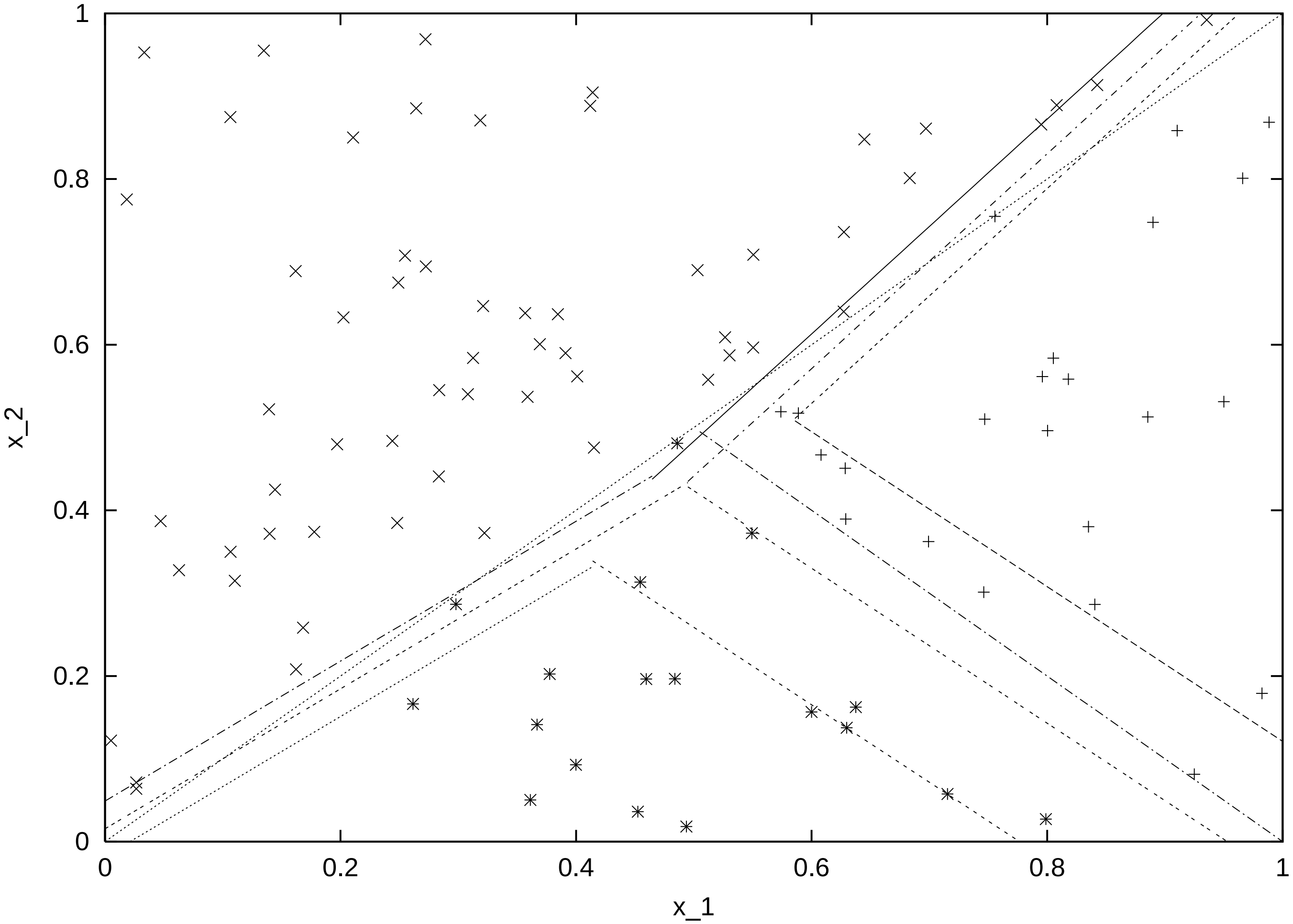

FIG. 1 – 3 categories linearly separable in 2D

Behaviour of a M-SVM (M-SVM1)

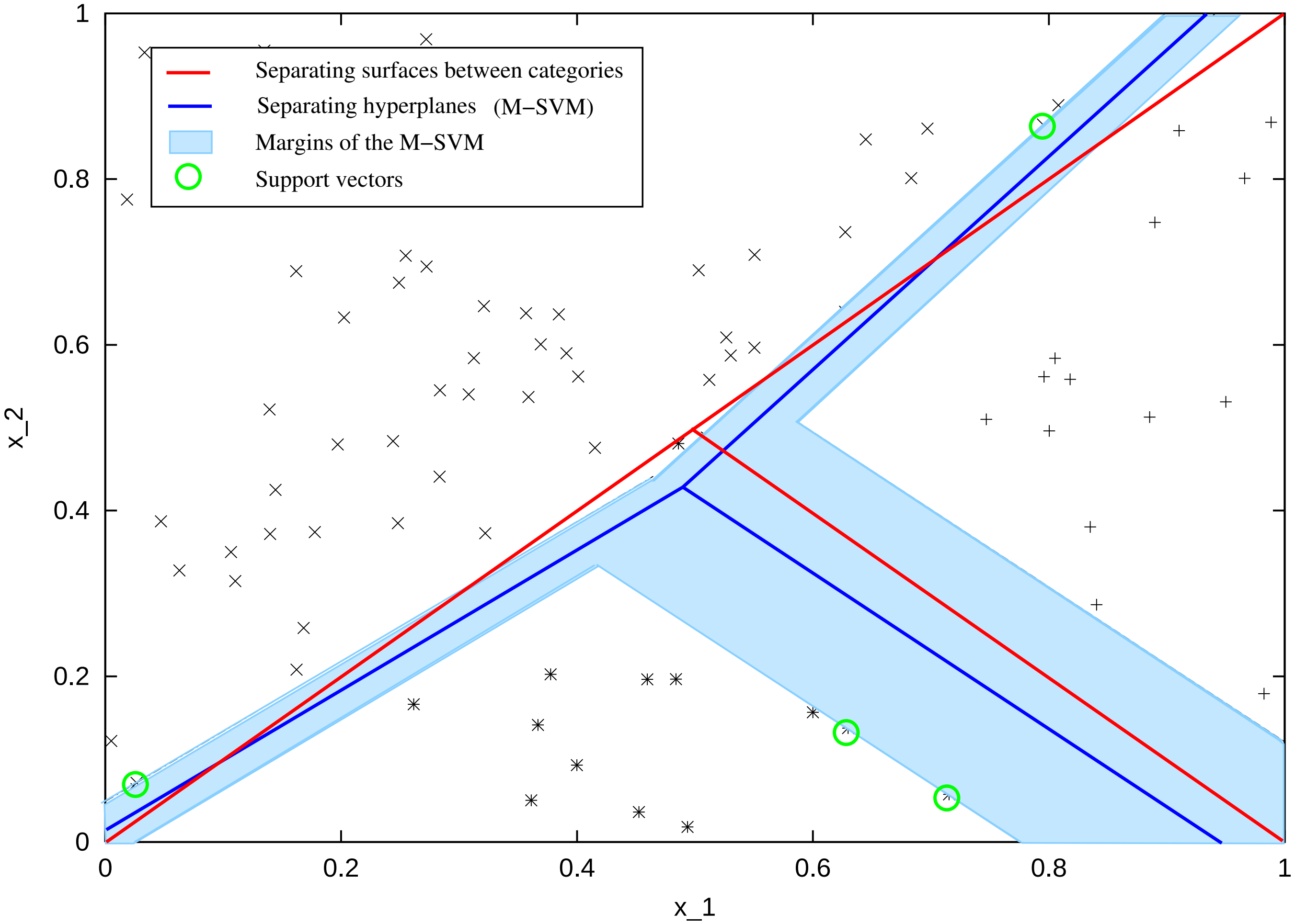

FIG. 2 – Hyperplanes of the linear M-SVM and support vectors

Behaviour of a M-SVM (M-SVM1)

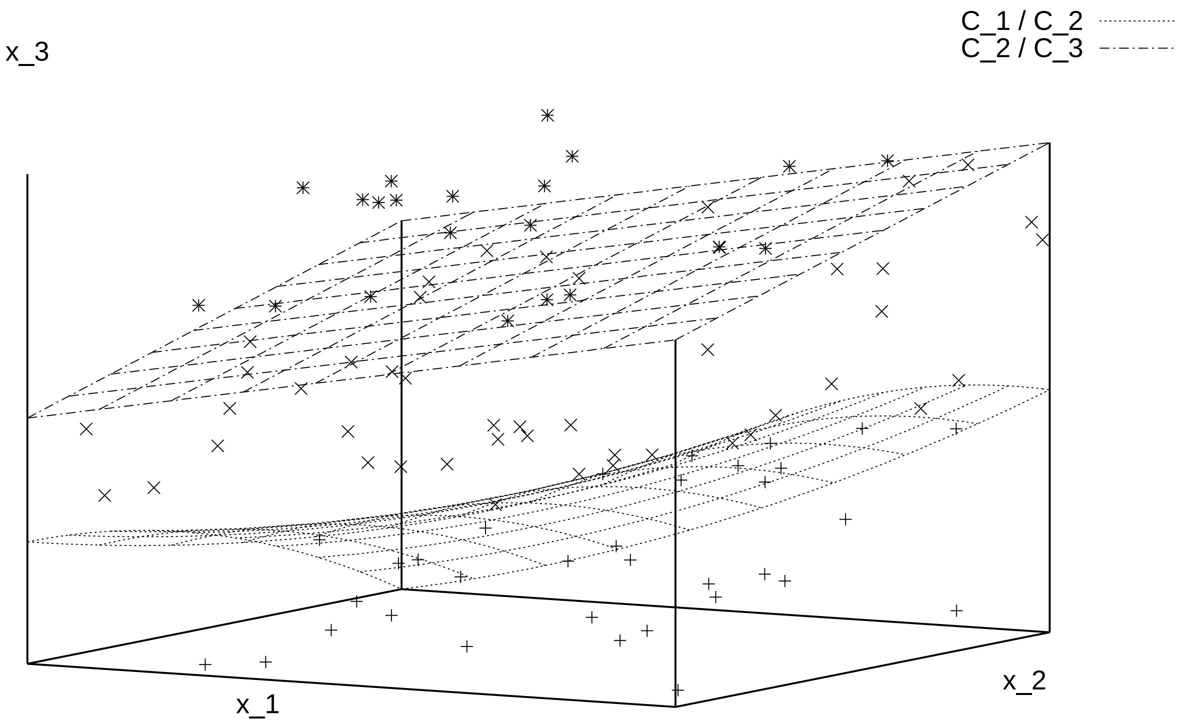

FIG. 3 – 3 categories nonlinearly separable in 3D

Behaviour of a M-SVM (M-SVM1)

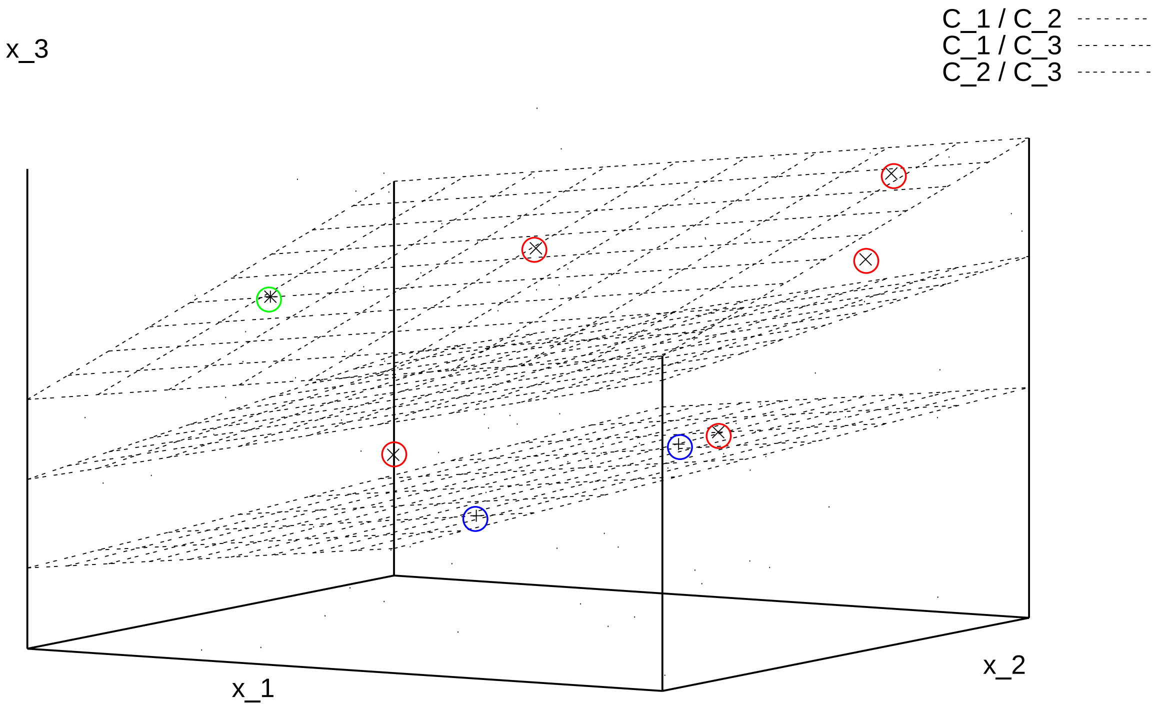

FIG. 4 – Hyperplanes of the linear M-SVM and support vectors

New kernel for protein sequence processing (Guermeur *et al.*, 03)

Analytical expression (primary structure only)

(x, x') : vectors encoding two window contents to be compared

$$\kappa_{\mu,\theta,A}(x, x') = \exp\left(-\mu \sum_{i=-n}^{n} \theta_i^2 (\|x_i\|^2 + \|x'_i\|^2 - 2\langle x_i, x'_i\rangle)\right)$$

Extension for multiple alignment processing

Straightforward : x replaced with $\tilde{x}$ such that $\tilde{x}_i = \sum_{j=1}^{22} \theta_{ij} a_j$

$$\langle \sum_{j=1}^{22} \theta_{ij} a_j, \sum_{k=1}^{22} \theta'_{ik} a_k \rangle = \sum_{j=1}^{22} \sum_{k=1}^{22} \theta_{ij} \theta'_{ik} \langle a_j, a_k \rangle$$

Taking into account the substitutions (matrix A)

Standard similarity matrices

G	2																			
P	1	3																		
D	0	0	2																	
E	0	-1	1	2																
A	0	-1	0	1	2															
N	0	0	1	0	0	3														
Q	0	0	0	1	0	1	2													
S	0	0	0	0	1	0	0	2												
T	0	0	0	0	0	0	0	0	2											
K	0	0	0	0	0	1	0	0	0	2										
R	0	0	0	0	0	0	0	0	0	1	2									
H	0	0	0	0	0	0	0	0	0	0	0	2								
V	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	2							
I	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	1	2						
M	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	0	0	2					
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2				
L	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	1	0	2	0	2			
F	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	0	1	0	-1	0	2		
Y	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0	-1	0	1	2	
W	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	0	0	0	0	-1	0	0	0	2
	G	**P**	**D**	**E**	**A**	**N**	**Q**	**S**	**T**	**K**	**R**	**H**	**V**	**I**	**M**	**C**	**L**	**F**	**Y**	**W**

FIG. 5 – Secondary structure similarity matrix **(Levin *et al.*, 86)**

Matrix M, characterizing the similarity between amino acids, is provided by the biologist

Approximating M with a Gram matrix

- $A = [a_1, \ldots, a_i, \ldots, a_{22}]^T \in \mathcal{M}_{22}(\mathbb{R})$: (implicit) representations of the amino acids
- AA^T : matrix of dot products = symmetric semi-definite positive approximation of M

Let the diagonalization of M be given by :

$$M = PDP^{-1} = PDP^T$$

(P is orthogonal since M is symmetric)

Then

$$AA^T = M_+ = PD_+P^T$$

where D_+ is derived from D by setting to 0 the negative eigenvalues

This leads to

$$A = P\sqrt{D_+}$$

Estimation of vector θ

Kernel target alignment : multi-class case (Vert, 02)

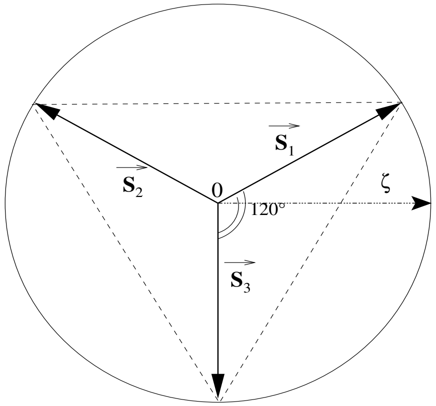

FIG. 6 – Optimal choice of the representatives of the categories (3-class case)

$$\kappa_t(x, x') = \begin{cases} \zeta^2 & \text{if } C(x) = C(x') \\ -\zeta^2/(Q-1) & \text{otherwise} \end{cases}$$

Training algorithm

Stochastic gradient descent

Estimation of the "width" μ

Free parameter of kernel target alignment

- Choosing ζ amounts to choosing θ up to a multiplicative constant
- $\Longrightarrow$ Choosing ζ amounts to choosing μ
- $\Longrightarrow$ ζ is set to an arbitrary value and μ is chosen in consequence (separately)

Alternative possibilities to set parameter μ

- Use resampling bounds devoted to M-SVMs (Y. Darcy)
- Use guaranteed risks for large-margin multi-class discrimimant models

This presentation deals with the second solution

Guaranteed risk - bi-class case - binary-valued functions

- $h = f$: takes its values in $\{-1, 1\}$
- $\mathcal{Y} = \{-1, 1\}$

Empirical risk

Definition 1 (Empirical risk) *Let f be a binary-valued function on $\mathcal{X}$. For a training data sequence $s_m = ((x_1, y_1), \ldots, (x_m, y_m))$ of length m,*

$$R_{s_m}(f) = \frac{1}{m} \left| \{(x_i, y_i) \in s_m : \; y_i f(x_i) = -1\} \right|$$

Growth function

Definition 2 (Growth function (Vapnik & Chervonenkis, 71)) *Let $\mathcal{F}$ be a set of indicator functions on $\mathcal{X}$. Let $\Pi_{\mathcal{F}}$ be the function which maps any set s of points in $\mathcal{X}$ to the number of dichotomies $\Pi_{\mathcal{F}}(s)$ computed on it by the functions in $\mathcal{F}$. The* growth function *of $\mathcal{F}$, $G_{\mathcal{F}}$, is the function from the nonnegative integers to the nonnegative integers given by :*

$$G_{\mathcal{F}}(m) = \max_{s_m \in \mathcal{X}^m} \Pi_{\mathcal{F}}(s_m)$$

Guaranteed risk - bi-class case - binary-valued functions

Vapnik-Chervonenkis (VC) dimension

Definition 3 (Vapnik & Chervonenkis, 71) *Let $\mathcal{F}$ be a set of indicator functions on a set $\mathcal{X}$. A subset s_m of $\mathcal{X}^m$ is said to be* shattered *by $\mathcal{F}$ if $\Pi_{\mathcal{F}}(s_m) = 2^m$, i.e. if each dichotomy on s_m is computed by a function of $\mathcal{F}$. The* VC dimension *of $\mathcal{F}$, VC-dim($\mathcal{F}$), is the largest value of m such that $G_{\mathcal{F}}(m) = 2^m$, if this value is finite, of infinity otherwise.*

If the VC dimension is finite, it is thus the size of the largest set of points shattered by $\mathcal{F}$.

Uniform convergence result

Theorem 1 (Vapnik, 82 $\rightarrow$)

$$\mathbb{P}\left\{\sup_{f \in \mathcal{F}} |R(f) - R_{s_m}(f)| > \delta\right\} < 4G_{\mathcal{F}}(2m) \exp\left(-m\left(\delta - \frac{1}{m}\right)^2\right)$$

Central importance of the VC dimension

Sauer's lemma

Theorem 2 (Sauer, 72, Shela, 72...) *If $\mathcal{F}$ has a finite VC dimension, then for $m \geq VC\text{-}dim(\mathcal{F})$, the following inequality holds true :*

$$G_{\mathcal{F}}(m) < \left(\frac{em}{VC\text{-}dim(\mathcal{F})}\right)^{VC\text{-}dim(\mathcal{F})}$$

VC dimension of a perceptron

Theorem 3 *Let $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{F}$ be the set of indicator functions of a linear (affine) separator in $\mathcal{X}$. Then :*

$$VC\text{-}dim(\mathcal{F}) = n + 1$$

Guaranteed risk - bi-class case - real-valued functions

- $h : \mathcal{X} \longrightarrow \mathbb{R}$
- $\mathcal{Y} = \{-1, 1\}$

Empirical margin risk

Definition 4 (Empirical margin risk (Bartlett, 98)) *Let h be a real-valued function on $\mathcal{X}$. For a training data sequence $s_m = ((x_1, y_1), \ldots, (x_m, y_m))$ of length m and a real number $\gamma > 0$*

$$R_{\gamma, s_m}(h) = \frac{1}{m} \left| \{(x_i, y_i) \in s_m : \; y_i h(x_i) < \gamma\} \right|$$

For $\gamma \in (0, 1]$, let $\pi_\gamma : \mathbb{R} \to [-\gamma, \gamma]$ be the piecewise-linear squashing function defined as

$$\pi_\gamma(x) = \begin{cases} \gamma.sign(x) & if \; |x| \geq \gamma \\ x & otherwise \end{cases}$$

Guaranteed risk - bi-class case - real-valued functions

Capacity measure : covering numbers

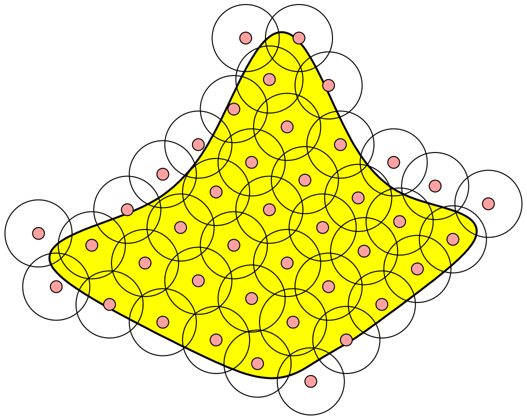

FIG. 7 – ϵ-net of a set $\mathcal{G}$ in a pseudo-metric or Banach space

Definition 5 (Covering numbers)
$\mathcal{N}(\epsilon, \mathcal{G}, \|.\|) =$ *minimum number of balls of radius* ϵ *required to cover the set* $\mathcal{G}$

Guaranteed risk - bi-class case - real-valued functions

Uniform convergence result

Theorem 4 (Bartlett, 98) *Let s_m be a m-sample of examples drawn independently from F. With probability at least $1-\delta$, for every value of γ in $(0,1]$, the risk $R(h)$ of a function h computed by a numerical bi-class discriminant model $\mathcal{H}$ is bounded above by :*

$$R(h) \leq R_{\gamma,s_m}(h) + \sqrt{\frac{2}{m}\left(\ln\left(2\mathcal{N}_\infty(\gamma/2, \mathcal{H}_\gamma, 2m)\right) + \ln\left(\frac{2}{\gamma\delta}\right)\right)}$$

where $\mathcal{H}_\gamma = \{\pi_\gamma \circ h : \; h \in \mathcal{H}\}$

$$\forall s_m \in \mathcal{X}^m, \; \forall (h^{(1)}, h^{(2)}) \in \mathcal{H}^2, \; d_{l_\infty(s_m)}(h^{(1)}, h^{(2)}) = \max_{x_i \in s_m} \left| h^{(1)}(x_i) - h^{(2)}(x_i) \right|$$

$$\mathcal{N}_\infty(\gamma/2, \mathcal{H}_\gamma, 2m) = \max_{s_{2m} \in \mathcal{X}^{2m}} \mathcal{N}(\gamma/2, \mathcal{H}_\gamma, d_{l_\infty(s_{2m})})$$

Guaranteed risk - bi-class case - real-valued functions

Fat-shattering dimension

Definition 6 (Fat-shattering dimension (Kearns & Schapire, 90)) *Let $\mathcal{H}$ be a set of real-valued functions on a set $\mathcal{X}$. For $\gamma > 0$, a subset $s_m = \{x_i\}$, $(1 \leq i \leq m)$ of $\mathcal{X}$ is said to be* γ-shattered *by $\mathcal{H}$ if there is a vector $v_b = [b_i] \in \mathbb{R}^m$ such that, for each binary vector $v_y = [y_i] \in \{-1, 1\}^m$, there is a function $h_y \in \mathcal{H}$ satisfying*

$$(h_y(x_i) - b_i)\, y_i \geq \gamma, \; (1 \leq i \leq m)$$

The vector v_b is then said to witness *the γ-shattering of s_m by $\mathcal{H}$. The* fat-shattering dimension *$fat_{\mathcal{H}}$ of the set $\mathcal{H}$ is a function from the positive real numbers to the integers which maps a value γ to the size of the largest set γ-shattered by functions of $\mathcal{H}$, if this size is finite, or to infinity otherwise.*

Central importance of the fat-shattering dimension

Extended Sauer's lemma

Theorem 5 (Alon *et al.*, 97) *Let $\mathcal{H}$ be a set of functions from $\mathcal{X}$ into $[0,1]$. For every value of γ in $(0,1]$ and every value of m in $\mathbb{N}^*$, the following bound is true :*

$$\mathcal{N}_\infty(\gamma, \mathcal{H}, m) \leq 2\left(\frac{4m}{\gamma^2}\right)^{d\log_2(2em/(d\gamma))}$$

where $d = fat_{\mathcal{H}}(\gamma/4)$.

Fat-shattering dimension of a perceptron

Theorem 6 (Bartlett & Shawe-Taylor, 99) *Suppose that $\mathcal{X}$ is the ball of radius $\Lambda_{\mathcal{X}}$ in a Hilbert space $E_{\mathcal{X}}$, and $\mathcal{H}$ is a linear separator computing the functions h given by : $\forall x \in \mathcal{X},\ h(x) = \langle w, x\rangle + b$, with $\|w\| \leq \Lambda_w$. Then :*

$$fat_{\mathcal{H}}(\epsilon) \leq \left(\frac{\Lambda_{\mathcal{X}}\Lambda_w}{\epsilon}\right)^2$$

Guaranteed risk - multi-class case - $\{1, \ldots, Q\}$-valued functions

- $h : \mathcal{X} \longrightarrow \{1, \ldots, Q\}$
- $\mathcal{Y} = \{1, \ldots, Q\}$

Ψ-dimension(s)

Definition 7 (Ψ-dimension (Ben-David *et al.*, 95)) *Let $\mathcal{F}$ be a set of functions on a set $\mathcal{X}$ taking their values in the finite set $\{1, \ldots, Q\}$. Let Ψ be a family of mappings ψ from $\{1, \ldots, Q\}$ into $\{-1, 1, *\}$, where $*$ is thought of as a null element. A subset $s_m = \{x_i\}$, $(1 \leq i \leq m)$ of $\mathcal{X}$ is said to be* Ψ-shattered *by $\mathcal{F}$ if there is a mapping $\psi^m = \left\{\psi^{(1)}, \ldots, \psi^{(m)}\right\}$ in Ψ^m such that for each vector v_y of $\{-1, 1\}^m$, there is a function f_y in $\mathcal{F}$ satisfying $\left[\psi^{(1)} \circ f_y(x_1), \ldots, \psi^{(i)} \circ f_y(x_i), \ldots, \psi^{(m)} \circ f_y(x_m)\right]^T = v_y$.*

Let $\mathcal{F}$ and Ψ be defined as above. The Ψ-dimension *of $\mathcal{F}$, denoted by Ψ-dim$(\mathcal{F})$, is the maximal cardinality of a subset of $\mathcal{X}$ Ψ-shattered by $\mathcal{F}$, if it is finite, or infinity otherwise.*

Guaranteed risk - multi-class case - $\{1, \ldots, Q\}$-valued functions

Main examples of Ψ-dimensions

Definition 8 (Graph dimension (Dudley, 87, Natarajan, 89)) *Let $\mathcal{F}$ be a set of functions from $\mathcal{X}$ into $\{1, \ldots, Q\}$. The* graph dimension *of $\mathcal{F}$, G-dim($\mathcal{F}$), is the Ψ-dimension of $\mathcal{F}$ in the specific case where Ψ is the set of Q mappings ψ_k, $(1 \leq k \leq Q)$, such that ψ_k takes the value 1 if its argument is equal to k, and the value -1 otherwise. Reformulated in the context of multi-class discriminant analysis, the functions ψ_k are the indicator functions of the categories.*

Definition 9 (Natarajan dimension (Natarajan, 89)) *Let $\mathcal{F}$ be a set of functions from $\mathcal{X}$ into $\{1, \ldots, Q\}$. The* Natarajan dimension *of $\mathcal{F}$ is the Ψ-dimension of $\mathcal{F}$ in the specific case where Ψ is the set of $\binom{Q}{2}$ mappings $\psi_{k,l}$, $(1 \leq k < l \leq Q)$, such that :*

$$\psi_{k,l}(t) = \begin{cases} 1 & \textit{if } t = k \\ -1 & \textit{if } t = l \\ * & \textit{otherwise} \end{cases}$$

Central importance of the graph dimension

Sauer's lemma

Directly applies in its standard formulation, since the graph dimension of $\mathcal{F}$ is simply the VC dimension of the *graph space* (set of graphs) of $\mathcal{F}$

Uniform convergence result

Theorem 7 (Anthony, 97) *Let $G_{G,\mathcal{F}}$ be the growth function of $\mathcal{F}$ associated with the graph dimension. Then :*

$$\mathbb{P}\{\exists f \in \mathcal{F} : \ (R_{s_m}(f) \leq (1-\delta)\epsilon \ \wedge \ R(f) > \epsilon)\} < 4G_{G,\mathcal{F}}(2m) \exp\left(-\frac{\delta^2 \epsilon m}{4}\right)$$

Guaranteed risk - multi-class case - vector-valued functions

- $h = [h_k] : \mathcal{X} \longrightarrow \mathbb{R}^Q$
- $\mathcal{Y} = \{1, \ldots, Q\}$

Definition 10 (Δ operators (Elisseeff *et al.*, 99 $\longrightarrow$)) *Let $h = [h_k]$, $(1 \leq k \leq Q)$ be a function from $\mathcal{X}$ into $\mathbb{R}^Q$.*

-Function $\Delta h = [\Delta h_k]$, $(1 \leq k \leq Q)$, from $\mathcal{X}$ into $\mathbb{R}^Q$, is given by :

$$\forall k \in \{1, \ldots, Q\}, \quad \Delta h_k(x) = \frac{1}{2} \min_{l \neq k} \{h_k(x) - h_l(x)\} = \frac{1}{2} \left\{ h_k(x) - \max_{l \neq k} h_l(x) \right\}$$

-Function $\Delta^* h = [\Delta^* h_k]$, $(1 \leq k \leq Q)$, from $\mathcal{X}$ into $\mathbb{R}^Q$, is given by :

Let $M(h, x) = \frac{1}{2} \max_k \{h_k(x) - \max_{l \neq k} h_l(x)\}$.

$$\forall k \in \{1, \ldots, Q\}, \quad \Delta^* h_k(x) = \begin{cases} M(h,x) & if \ \frac{1}{2}\{h_k(x) - \max_{l \neq k} h_l(x)\} = M(h,x) \\ -M(h,x) & otherwise \end{cases}$$

Guaranteed risk - multi-class case - vector-valued functions

Empirical margin risk

Definition 11 (Empirical margin risk (Elisseeff *et al.*, 99)) *The empirical risk with margin* $\gamma \in (0, 1]$ *of* h *on a set* $s_m = \{(x_1, C(x_1)), \ldots, (x_m, C(x_m))\}$ *of size* m *is :*

$$\begin{aligned} R_{\gamma, s_m}(h) &= \frac{1}{m} \left| \left\{ (x_i, C(x_i)) \in s_m : \ \Delta h_{C(x_i)}(x_i) < \gamma \right\} \right| \\ &= \frac{1}{m} \left| \left\{ (x_i, C(x_i)) \in s_m : \ \Delta^* h_{C(x_i)}(x_i) < \gamma \right\} \right| \end{aligned}$$

Basic capacity measure

$\Delta_\gamma h = [\pi_\gamma \circ \Delta h_k]$, $(1 \leq k \leq Q)$, $\Delta_\gamma \mathcal{H} = \{\Delta_\gamma h : \ h \in \mathcal{H}\}$, $\Delta^*_\gamma h = [\pi_\gamma \circ \Delta^* h_k] \ldots$

$$\forall s_m \in \mathcal{X}^m, \ \forall (h^{(1)}, h^{(2)}) \in \mathcal{H}^2, \ d_{l_\infty, l_\infty(s_m)}(h^{(1)}, h^{(2)}) = \max_{x_i \in s_m} \max_{k \in \{1, \ldots, Q\}} \left| h_k^{(1)}(x_i) - h_k^{(2)}(x_i) \right|$$

$$\mathcal{N}_{\infty,\infty}(\gamma, \mathcal{H}, m) = \max_{s_m \in \mathcal{X}^m} \mathcal{N}(\gamma, \mathcal{H}, d_{l_\infty, l_\infty(s_m)})$$

Guaranteed risk - multi-class case - vector-valued functions

Uniform convergence result

Theorem 8 (Guermeur *et al.*, 04) *Let s_m be a m-sample of examples drawn independently from F. With probability at least $1-\delta$, for every value of γ in $(0,1]$, the risk $R(h)$ of a function h computed by a numerical Q-class discriminant model $\mathcal{H}$ is bounded above by :*

$$
\begin{aligned}
R(h) &\leq R_{\gamma, s_m}(h) + \sqrt{\frac{2}{m}\left(\ln\left(2\mathcal{N}_{\infty,\infty}(\gamma/4, \Delta_\gamma \mathcal{H}, 2m)\right) + \ln\left(\frac{2}{\gamma\delta}\right)\right)} + \frac{1}{m} \\
R(h) &\leq R_{\gamma, s_m}(h) + \sqrt{\frac{2}{m}\left(\ln\left(2\mathcal{N}_{\infty,\infty}(\gamma/4, \Delta^*_\gamma \mathcal{H}, 2m)\right) + \ln\left(\frac{2}{\gamma\delta}\right)\right)} + \frac{1}{m}
\end{aligned}
$$

Simple pathway to bound the covering numbers of M-SVMs (Williamson *et al.*, 01 $\longrightarrow$)

Theorem 9 *Let* $\mathcal{H}$ *be a set of functions from* $\mathcal{X}$ *into* $\mathbb{R}^Q$*. For every value of* ϵ *and* γ *satisfying* $0 < \epsilon < \gamma$*, the following bounds hold true :*

$$\mathcal{N}_{\infty,\infty}(\epsilon, \Delta_\gamma \mathcal{H}, m) \leq \mathcal{N}_{\infty,\infty}(\epsilon, \Delta \mathcal{H}, m) \leq \mathcal{N}_{\infty,\infty}(\epsilon, \mathcal{H}, m)$$

Theorem 10 *Let* $\mathcal{H}$ *be a set of functions computed by a M-SVM. Suppose that the biases* b_k *all belong to the interval* $[-b, b]$ *Then :*

$$\mathcal{N}_{\infty,\infty}(\epsilon, \mathcal{H}, m) \leq \left\lceil \frac{b}{\epsilon} \right\rceil^Q \mathcal{N}_{\infty,\infty}(2\epsilon, \tilde{\mathcal{H}}, m)$$

Covering numbers and entropy numbers

Definition 12 (Entropy numbers) *Let (E, ρ) be a pseudo-metric space and H a subset of E. Then, for n in $\mathbb{N}^*$, the n^{th} entropy number of H, $\epsilon_n(H)$, is :*

$$\epsilon_n(H) = \inf \{\epsilon > 0 : \ \mathcal{N}(\epsilon, H, \rho) \leq n\}$$

Definition 13 (Entropy numbers of a bounded linear operator) *Let E_H and F_H be Hilbert spaces endowed with the norms $\|.\|_{E_H}$ and $\|.\|_{F_H}$. Let $\mathcal{L}(E_H, F_H)$ be the set of all bounded linear operators between $(E_H, \|.\|_{E_H})$ and $(F_H, \|.\|_{F_H})$. Let $T \in \mathcal{L}(E_H, F_H)$.*

$$\|T\| = \sup_{x \in U_{E_H}} \|Tx\|_{F_H} \quad \epsilon_n(T) = \epsilon_n(T(U_{E_H}))$$

Theorem 11 (Maurey) *Let $T \in \mathcal{L}(E_H, \mathbb{R}^m)$. Then, there exists a constant c such that, for all $n \in \mathbb{N}^*$,*

$$\epsilon_n(T) \leq c\|T\| \left((\log(n) + 1)^{-1} \log \left(1 + \frac{m}{\log(n) + 1} \right) \right)^{1/2}$$

Covering numbers and entropy numbers

Theorem 12 *Let* $\tilde{\mathcal{H}} = \left\{ \tilde{h} \right\}$ *be defined as above, with the additional constraint that each function* $\tilde{h} = [w_k]$, $(1 \leq k \leq Q)$, *satisfies :* $\max_k \|w_k\|_2 \leq \Lambda_w$

Let T_{s_m} *be the linear operator given by* $T_{s_m} = \Lambda_w S_{s_m}$ *with :*

$$S_{s_m} : \begin{array}{l} l_2 \longrightarrow l_\infty^{Qm} \\ w \mapsto [\langle w_1, \Phi(x_1)\rangle, \dots \langle w_Q, \Phi(x_1)\rangle, \dots, \langle w_1, \Phi(x_m)\rangle, \dots \langle w_Q, \Phi(x_m)\rangle] \end{array}$$

Then

$$\epsilon_n\left(T_{s_m}\right) \leq \epsilon_0 \Longrightarrow \mathcal{N}_{\infty,\infty}(\epsilon_0, \tilde{\mathcal{H}}, m) \leq n$$

Scale-sensitive Ψ-dimensions

Definition 14 (γ-Ψ-shattering) *Let $\mathcal{H}$ be a set of functions on a set $\mathcal{X}$ taking their values in $\mathbb{R}^Q$. Let Ψ be a family of mappings ψ from $\{1, \ldots, Q\}$ into $\{-1, 1, *\}$. For $\gamma > 0$, a subset $s_m = \{x_i\}$, $(1 \leq i \leq m)$ of $\mathcal{X}$ is said to be* γ-Ψ-shattered *by $\mathcal{H}$ if there is a mapping $\psi^m = \left(\psi^{(1)}, \ldots, \psi^{(i)}, \ldots, \psi^{(m)}\right)$ in Ψ^m and a vector $v_b = [b_i]$ in $\mathbb{R}^m$ such that, for each vector $v_y = [y_i]$ of $\{-1, 1\}^m$, there is a function h_y in $\mathcal{H}$ satisfying*

$$\forall i \in \{1, \ldots, m\}, \left\{ \begin{array}{lll} \text{if } y_i = & 1 & \exists (k,l) : \left\{ \begin{array}{rcrcr} (\psi^{(i)}(k) & = & 1 & \wedge & h_{y,k}(x_i) - b_i \geq \gamma) \\ (\psi^{(i)}(l) & = & -1 & \wedge & h_{y,l}(x_i) = -h_{y,k}(x_i)) \end{array} \right. \\ \text{if } y_i = & -1 & \exists (k,l) : \left\{ \begin{array}{rcrcr} (\psi^{(i)}(k) & = & 1 & \wedge & h_{y,k}(x_i) - b_i \leq -\gamma) \\ (\psi^{(i)}(l) & = & -1 & \wedge & h_{y,l}(x_i) = -h_{y,k}(x_i)) \end{array} \right. \end{array} \right.$$

Definition 15 (Ψ-dimension with margin γ) *Let $\mathcal{H}$, Ψ and γ be defined as above. The Ψ-dimension of $\mathcal{H}$ with margin γ, denoted by Ψ-dim$(\mathcal{H}, \gamma)$, is the maximal cardinality of a subset of $\mathcal{X}$ γ-Ψ-shattered by $\mathcal{H}$, if it is finite, or infinity otherwise.*

Margin Natarajan dimension

Definition 16 (Natarajan dimension with margin γ) *Let $\mathcal{H}$ be a set of functions on a set $\mathcal{X}$ taking their values in $\mathbb{R}^Q$. For $\gamma > 0$, a subset $s_m = \{x_i\}$, $(1 \leq i \leq m)$ of $\mathcal{X}$ is said to be* γ-N-shattered *(N-shattered with margin γ) by $\mathcal{H}$ if there is a set of couples of distinct indexes*

$$I(s_m) = \left((i_1(x_1), i_2(x_1)), \ldots, (i_1(x_i), i_2(x_i)), \ldots, (i_1(x_m), i_2(x_m))\right)$$

in $(\{1, \ldots, Q\} \times \{1, \ldots, Q\})^m$ and a vector $v_b = [b_i]$ in $\mathbb{R}^m$ such that, for each binary vector $v_y = [y_i] \in \{-1, 1\}^m$, there is a function h_y in $\mathcal{H}$ satisfying

$$\forall i \in \{1, \ldots, m\}, \begin{cases} \textit{if } y_i = 1 \textit{ then} \left(h_{y,i_1(x_i)}(x_i) - b_i \geq \gamma \wedge h_{y,i_2(x_i)}(x_i) = -h_{y,i_1(x_i)}(x_i)\right) \\ \textit{if } y_i = -1 \textit{ then} \left(h_{y,i_1(x_i)}(x_i) - b_i \leq -\gamma \wedge h_{y,i_2(x_i)}(x_i) = -h_{y,i_1(x_i)}(x_i)\right) \end{cases}$$

The Natarajan dimension with margin γ *of the class $\mathcal{H}$, N-dim$(\mathcal{H}, \gamma)$, is the maximal cardinality of a subset of $\mathcal{X}$ γ-N-shattered by $\mathcal{H}$, if it is finite, or infinity otherwise.*

Central importance of the margin Natarajan dimension

Extended Sauer's lemma

Theorem 13 *Let* $\mathcal{H}$ *be a class of functions from a domain* $\mathcal{X}$ *into* $\mathbb{R}^Q$*. For every value of* γ *in* $(0,1]$ *and every integer value of* m *satisfying* $2m \geq N\text{-}dim\left(\Delta_\gamma^* \mathcal{H}, \gamma/48\right)$*, the following bound is true :*

$$\mathcal{N}_{\infty,\infty}(\gamma/4, \Delta_\gamma^* \mathcal{H}, 2m) < 2\left(98mQ(Q-1)\right)^{d \log_2(98emQ(Q-1)/d)}$$

where $d = N\text{-}dim\left(\Delta_\gamma^* \mathcal{H}, \gamma/48\right)$*.*

Margin Natarajan dimension of the M-SVMs

Theorem 14 *Let $\tilde{\mathcal{H}}$ be the class of functions computed by a Q-class linear (without bias) SVM on a domain $\mathcal{X}$. Suppose that $\Phi(\mathcal{X})$ is included in the ball of radius $\Lambda_{\Phi(\mathcal{X})}$ in $E_{\Phi(\mathcal{X})}$ and the vectors w_k defining the hyperplanes satisfy $1/2 \max_{1 \leq k < l \leq Q} \|w_k - w_l\| \leq \Lambda_w$. Then, for any positive real value ϵ, the following bound holds true :*

$$N\text{-}dim\left(\Delta\tilde{\mathcal{H}}, \epsilon\right) \leq \binom{Q}{2} \left(\frac{\Lambda_w \Lambda_{\Phi(\mathcal{X})}}{\epsilon}\right)^2$$

Guaranteed risk - multi-class case - vector-valued functions

Open problems (short term)

- Derive a generalized Sauer's lemma for the margin Natarajan dimension of $\Delta_\gamma \mathcal{H}$
- Bound the *margin graph dimension*
- Use these bounds efficiently to perform a line search on the width of the kernel

Conclusions and future work

Conclusions

- Dedicated solutions are needed to tune kernel parameters in the multi-class case
- Part of the standard results can be extended in a straightforward way... but not all!
- New tools/approaches are needed to derive tight bounds on the expected risk

Future work

- Extension of the bounds based on the leave-one-out procedure (Y. Darcy)
- Study of alternative measures of complexity (Donsker classes...) (O. Teytaud)
- Use of concentration inequalities (S. Boucheron, R. Vert,...)